Thesis for the Degree of Licentiate of Engineering

Statistical methods for early discovery of diabetic neuropathy using epidermal nerve fiber data

Claes Andersson

CHALMERS | UNIVERSITY OF GOTHENBURG

Division of Mathematical Statistics
Department of Mathematical Sciences
Chalmers University of Technology
and University of Gothenburg
Gothenburg, Sweden, 2016

Statistical methods
for early discovery of diabetic neuropathy using epidermal nerve fiber data
Claes Andersson

Department of Mathematical Sciences
Chalmers University of Technology
and University of Gothenburg
SE–412 96 Gothenburg
Sweden
Phone: +46 (0)31-772 10 00

Printed in Gothenburg, Sweden, 2016.

Statistical methods for early discovery of diabetic neuropathy using epidermal nerve fiber data

Claes Andersson

Department of Mathematical Sciences
Chalmers University of Technology
and
University of Gothenburg

Abstract

The main aim with the work in this thesis is to increase the understanding of the effects diabetic neuropathy has on the epidermal nerve fibers and thereby find methods to detect the disorder at an earlier stage. Epidermal nerve fibers (ENFs) are small sensory nerve fibers in the skin, sensing heat and pain. Earlier diagnosis of the disorder can help to slow down the progression and delay the symptoms. The data used are skin samples from a group of 32 healthy volunteers and 20 diabetic subjects with differently progressed diabetic neuropathy, in which the nerve fibers have been traced using confocal microscopy. One part of the work is based on methods from spatial statistics, considering the points where the nerve fibers enter the epidermis and where they terminate as realizations of point processes. The point patterns obtained from healthy subjects are compared to those of subjects at an early stage of the neuropathy, in terms of spatial summary statistics, including a new tool to quantify the area of the skin covered by a nerve, the *reactive territory*. Significant differences between the groups are found, that has not previously been reported. Moreover, a point process model for the nerve fiber patterns is proposed, to help the understanding of the growth process of the nerve fibers. In the other part of the work, hierarchical models for the nerve fiber segments are proposed, and used to perform unsupervised classification. The results are evaluated in terms of how well the diabetic subjects are separated from the healthy ones. It is found that the results are considerably improved when including the information about the nerve fiber segments, compared to only using the number of nerve fibers.

Keywords Diabetic neuropathy, Epidermal nerve fibers, Hierarchical models, Point processes, Reactive territory.

List of appended papers

The following papers are included in this thesis.

A. Discovering early diabetic neuropathy from epidermal nerve fiber patterns. Claes Andersson, Peter Guttorp, Aila Särkkä. (submitted)

B. Hierarchical models for epidermal nerve fiber patterns. Claes Andersson, Tuomas Rajala, Aila Särkkä. (preprint)

Acknowledgements

First and foremost, I would like to thank my supervisor Aila Särkkä, for her constant encouragement, guidance and support. I am very grateful for the fact that you always find the time to discuss my work, despite your many other responsibilities, and for making me feel more like a collaborator than a student. I would also like to thank Tuomas Rajala, my co-supervisor, for his many ideas and suggestions, without which this thesis would not be what it is, and for teaching me many tricks of the trade. Furthermore, I want to thank Peter Guttorp, coauthor of Paper A. It has been a great experience working with you.

A big thanks to my colleagues at the department, for making it the lovely workplace that it is. Especially Ivar, for being such a good friend, Malin, for your timely disturbances while I work, Peter and Dawan for your constant optimism, Magnus, for your constant pessimism, and Henrike, for having answers to all the questions an academic little brother might have.

Finally, I want to thank my family for all the love and support. Thank you for always being there, without you I would not be where I am.

Contents

Chapter One

Introduction

Epidermal nerve fibers (ENFs) are thin sensory nerve fibers in the epidermis, the outermost layer of the skin, that transmit signals of heat and pain. The ENFs enter the epidermis as single nerve fibers cross the junction between the epidermis and the dermis (the layer below the epidermis), and extend into the epidermis, with or without branching, before terminating. The existence of ENFs was theorized for over 130 years, before it was conclusively established by William Kennedy and Gwen Wendelschafer-Crabb, using confocal microscopy (Kennedy and Wendelschafer-Crabb, 1993). Figure 1.1 shows a side section of a skin blister with the nerve fibers visible (left panel). Also in Figure 1.1 are illustrations of the types of data used in this work. In the top right panel are ENF data as a point pattern, where the points are the locations where the nerves enter the epidermis and where they terminate, and in the bottom right panel is part of the same sample with the branching points added (right panel). Henceforth, the points where nerve fibers enter the epidermis are referred to as base points and the points where they terminate as end points.

Once the methods for visualizing and identifying ENFs were established, research moved towards quantification of such nerve fibers, in particular to assess their potential use in diagnosing peripheral neuropathy, which is damage or disease affecting the nerves. Neuropathy caused by diabetes is called diabetic neuropathy. The symptoms include loss of sensation and neuropathic pain, and can drastically reduce the life quality of an affected subject. Although there is no treatment to cure a person from diabetic neuropathy, measures can be taken to slow down the progression of the disease. Therefore it is of interest to be able to detect it as early as possible.

It is well established that neuropathy causes the number of nerves to decrease, and the use of ENF data in current clinical practice is largely based on ENF density, i.e. the observed number of nerves in the epidermis (Lauria et al., 2010b). A large number of studies have been dedicated to establish a normative reference range for the ENF density. One of the larger studies included eight laboratories world wide, and established different age dependent normative reference ranges for men and women (Lauria et al., 2010a). However, it has also been noted that the nerve fibers in subjects with diabetic neuropathy exhibit a more clustered pattern than in healthy subjects (Kennedy et al., 1999). This

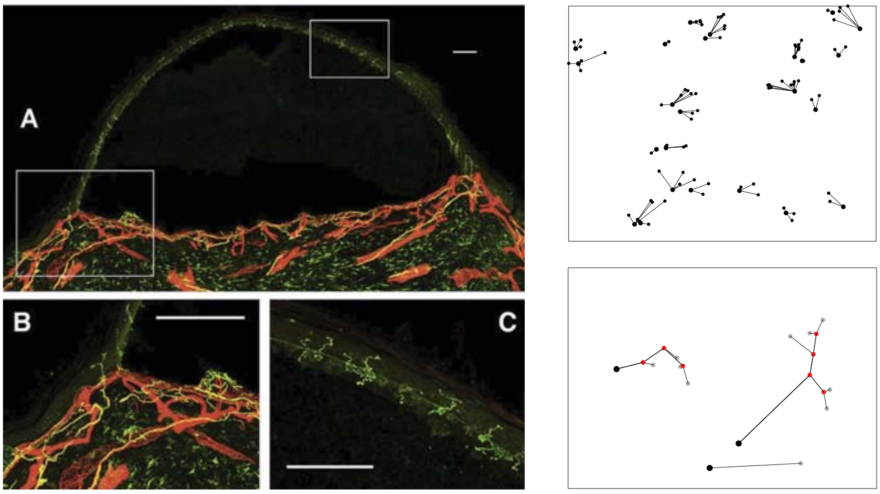

Figure 1.1: Left panel: A side view of a suction induced skin blister, with the ENFs visible. The nerve fibers are bright green to yellow, while blood vessels are red. Samples are collected as parts of the blister roof seen in Figure A. Figure B shows a dermal papilla, a ridge in the dermis, typically with an aggregation of nerve fibers. In Figure C the ENFs are zoomed in, with the nerve fibers clearly visible. Top right panel: ENF data illustrated as a point pattern. Large dots are base points and smaller dots are end points. Bottom right panel: part of the same sample with the base points as black dots, branching points as red dots and end points as gray dots.

observation has been quantified in several studies, using methods from spatial statistics. Using samples from the thigh of one healthy subject, and the thighs of two subjects diagnosed with mild, two with moderate and two with severe diabetic neuropathy, Waller et al. (2011) found that the patterns from the subjects with moderate and severe diabetic neuropathy were significantly more clustered than the patterns from the healthy subject, in terms of second order summary statistics. Furthermore, Myllymäki et al. (2014) included disease status (healthy or mild/moderate) as a covariate in a Gaussian process regression to model the second order summary statistic (Myllymäki et al., 2014). Also in this study the patterns from the diseased subjects were found to be clearly more clustered.

In addition, Myllymäki et al. (2012) focused on the effect on the spatial pattern of three non spatial covariates (age, gender and BMI), analyzing data from 25 healthy volunteers, and found that the covariates seem to affect the spatial pattern of the base points in the calf, while no significant effects on the patterns in the samples from the foot were found. Moreover, a point process model for the base and end point patterns of healthy volunteers, the non orphan cluster (NOC) model, was proposed by Olsbo et al. (2013).

The work in this thesis mainly aims at finding ways to detect the neuropathy at an early stage using the ENF data. It is a continuation of the work above, as tools from spatial analysis are part of the analysis. However, we also use another approach, not considering the spatial aspect, but rather focusing on the characteristics of individual nerve fibers. The results from this approach suggest that there are some differences between the healthy subjects and subjects diagnosed with diabetic neuropathy. Moreover, the differences characterize the way the neuropathy affects the individual nerve fibers, which has not been under much consideration before. In addition, our results offer a partial explanation of the results in earlier studies.

In the two papers included in this thesis we use various statistical methods to characterize the nerve fiber patterns and distinguish the patterns from the healthy subjects from the patterns from subjects suffering from diabetic neuropathy. In Paper A a point process approach is used to compare samples from the right foot of 32 healthy volunteers to similar samples from 8 subjects diagnosed with mild diabetic neuropathy. The base point and end point patterns are treated as realizations of point processes and several summary statistics for point patterns are used to compare the groups. In the analysis the information of which base point each end point is connected to is used, for example, to compare the distances between the base points and end points in the two groups. Moreover, a tool for quantifying the area of the skin covered by a nerve fiber is introduced, called the *reactive territory*, as well as a point process model for the end points conditional on a base point pattern. The new model is compared to the non orphan cluster (NOC) model introduced by Olsbo et al. (2013). The data analysis reveals several significant differences between the two groups, one of which is that the nerves in the mild group cover less of the skin in terms of their reactive territories. From the modelling it is clear that there are dependencies between the different nerve fibers of a more complex nature than what either model can capture. For the end points connected to the same base point, however, the new model seems to fit the data better than the NOC model does.

In Paper B the individual nerve fibers are in focus, which are studied in terms of their tree structure which is composed of the segments between the points where the nerve starts, branches and ends. Summaries of the data both for the nerve fibers as a whole and for the individual line segments are introduced, and some initial exploratory analysis of the data from 32 healthy volunteers and 20 subjects diagnoses with mild, moderate or severe diabetic neuropathy indicate that there are some differences between the disease groups. The main question, however, is if the subjects diagnosed with diabetic neuropathy can be separated from the healthy subjects based on such nerve tree data alone. For this, three models for the data are constructed, which are used to perform unsupervised classification. These models are compared to a baseline model that uses only the nerve counts from each subject, and the results indicate that the additional information in the tree structures improves the classification.

1.1 Data

The data we use consist of skin samples taken from 32 healthy volunteers and 20 diabetic subjects. The samples are obtained using suction induced skin biopsies, where a portion of the epidermis is removed, mounted on a slide and stained for imaging. The nerves are then traced using confocal microscopy, and the locations of the points where the nerve enter the epidermis, branch and terminate are recorded. The data are in 3D, and the samples are boxes of size $330\mu m \times 432\mu m \times z$, where z varies from 20 to 50 μm. From each subject, two blisters were taken from six different body parts. The part of the data we have consists of the blister specimens taken from the right foot and right calf of each subject. Since the effects of the neuropathy tend to appear the earliest in the most distal parts of the body, we mainly focus on the samples taken from the foot. From each subject, three to six images were taken, typically two from each blister. Moreover, covariates such as age, gender and BMI were recorded for each subject, as well as a neuropathy score (with higher score indicating more severe neuropathy).

In Paper A the 2D projections of the x- and y-coordinates are used, and the data are treated as point patterns considering only the base points and end points. In Paper B the 3D data are used, and the segments connecting the base, branching and end points in each nerve are used as observations.

1.2 Methods from spatial statistics

As mentioned earlier, one way to analyze the ENF data is to treat the data as a collection of point patterns, where only the information about the coordinates where nerve fibers enter the epidermis and where they terminate is used. When analysing point patterns, a commonly occurring question is whether the points can be considered completely spatially random or if the underlying mechanism generating the patterns gives some structure to them. In Figure 1.2 three point patterns are visualized. The middle pattern is a realization of the case when the points are independently and uniformly scattered in the observation window, commonly referred to as complete spatial randomness (CSR). This means that the information that there is a point of the process at a certain location does not effect the probability of finding a point of the process anywhere else (in the observation window). Such a process is referred to as a homogeneous Poisson process, which can be defined through two properties. Firstly, that there is a constant $\lambda > 0$ such that for any bounded set, B, the number of points in B

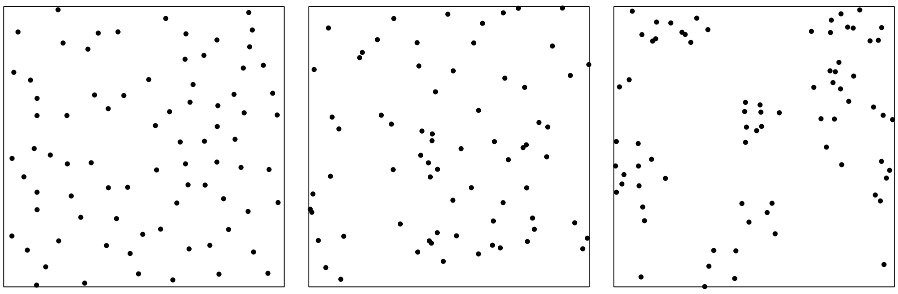

Figure 1.2: Realizations of three different point processes. The leftmost pattern is a realization of a process that exhibits regularity, the pattern in the center is a realization of a completely spatially random (CSR) process while the rightmost pattern is from a cluster process.

follows a Poisson distribution with parameter $\lambda|B|$. Secondly, given that the number of points in B is n, the n points are a random sample from the uniform distribution on B. The leftmost pattern, on the other hand, exhibits *regularity*. This means that the points of the pattern have a tendency to be further apart, compared to CSR. In other words, given the location of a point in the pattern, the probability of finding another point in its vicinity is lower than under CSR. The rightmost pattern of Figure 1.2 is an example of the opposite, where the points tend to appear in groups, which is called *clustering*.

The concepts of CSR, regularity and clustering are often used to characterize observed point patterns and the underlying process generating the patterns. In the following a brief introduction to the theory of point processes is given, leading up to a description of a few of the methods that can be used to analyze point patterns.

Point processes

Point processes have been used to study a wide variety of phenomena, including the locations of galaxies, trees in forest stands, ants' nests – and nerve fibers. The plots in Figure 1.2 show realizations of point processes, i.e. the patterns are to the respective point process what a single number is to a random variable describing the roll of a die or what "heads" or "tails" is to a coin flip. The realizations of a point process are clearly more complex than those of the random variables mentioned, and so is the theory to define and characterize the point processes, as one might expect. Only a fraction of the theory will be brought up here, and described with more focus on intuition than mathematical rigour. More extensive and rigorous treatments of the subject can be found in e.g. Illian et al. (2008); Møller and Waagepetersen (2003); Chiu et al. (2013); Diggle (2014); Lieshout (2000); Gelfand et al. (2010), and Cressie (2015). The sections below mainly follow Illian et al. (2008).

Each of the patterns in Figure 1.2 can mathematically be described as an unordered set of points, $\mathbf{x}$, i.e

$$\mathbf{x} = \{x_1, \ldots, x_n\}, \tag{1.1}$$

where x_i describes the location of an observed point. Such an observation, $\mathbf{x}$, is commonly referred to as a *configuration.* Thus a point process is a stochastic mechanism, the realizations of which are random configurations of points in some space D. Point processes can be defined on more general topological spaces, but here the attention will be restricted to letting $D = \mathbb{R}^2$. Thus, a point process, X, in this setting may be defined as a random set of points

$$X = \{x_1, x_2, \ldots\} \tag{1.2}$$

where the number of points can be random or deterministic and where the locations, x_i, are random points in $\mathbb{R}^2$. Another way to define a point process is to view it as a random counting measure, that is, for any set $B \subset \mathbb{R}^2$ the point process maps the set to the random number of points of X falling in B. Here, this is used as a way to characterize the process, and $N_X(B)$ is the notation for the random number of points falling in the set B. A point process is *simple* if it places at most one point at any location and *locally finite* if for any bounded set B, the random variable $N_X(B)$ is finite.

Through the random variables $N_X(B)$ the *intensity measure* of a point process can be defined as $\Lambda(B) = \mathbb{E}(N_X(B))$, i.e. $\Lambda(B)$ is the expected number of points in B. Under some continuity conditions, which as usually satisfied in practical applications, Λ admits a density with respect to the Lebesgue measure, that is

$$\Lambda(B) = \int_B \lambda(x)dx \tag{1.3}$$

where $\lambda(x)$ is the *intensity function* of the point process.

One often considers point processes that are *stationary* and *isotropic*. A point process is stationary if its distribution is translation invariant, and isotropic if its distribution is invariant under rotations around the origin. Note that for a point process to be stationary, it must be defined on the whole of $\mathbb{R}^2$. In practice, however, the observations of a point process are confined to some bounded region, which is only a part of the area on which the process operates. This is the case, for example, when observing a tree stand that is part of a larger forest or, as in the data used in this work, where the samples are very small compared to the whole skin. Although the full area of the process is also most commonly bounded, as is the case in the examples above, the point process is usually defined on $D = \mathbb{R}^2$. The underlying assumption is that the distance from the boundary of the full area of the process to the observation region is large enough for any boundary effects to be ignored in the observed region. One important property of a stationary process is that its intensity function is constant, i.e. $\lambda(x) \equiv \lambda$ for all x, which implies that $\mathbb{E}(N(B)) = \lambda|B|$, where $|B|$ is the area of B.

So far point processes have been described as models for the locations of points, but one could include more information by attaching *marks* to the points. Using the tree stand example, the marks of the points could be the diameter of the trees. For the nerve fiber data marks indicating whether a point is a base point or an end point are used. The class of marked point process models is rich, as is the literature on the topic. Some of the earlier work can be found in Ogata and Tanemura (1985) and Takacs and Fiksel (1986), while Diggle (2014) and Illian et al. (2008) provide more recent examples. Since the use of marked point processes in this thesis is rather limited, the subject will not be treated further here.

Summary statistics

When analyzing point patterns, the aim is typically to reach an understanding of the process that generated the observed patterns. In this section some of the functional summary statistics that can be used for this purpose are introduced. They are all tools to describe the structure of the process, i.e. if the process generates clustered, regular or CSR patterns. It will be assumed that the process considered is stationary and isotropic, although there are versions of some of the summary statistics for non stationary processes as well. It should be noted, however, that unless some information that can explain differences in the intensity is available, there is no obvious way to differentiate between non stationarity and, for example, clustering. The way in which the summary statistics are used, is typically that they are estimated from data and compared to estimates from simulations under CSR or under some other null model. Some aspects of estimating these summary statistics from data are treated in the Edge correction section.

One way to describe the structure of a point process is through the *empty space function*, $F(r)$, which gives the probability that the distance from a random location, x, to the nearest point of the process is less than or equal to r. As the process is assumed to be stationary, the random location can be replaced by the origin. Thus, this can be formulated as

$$F(r) = 1 - P(N_X(b(o,r)) = 0), \tag{1.4}$$

where $b(o,r)$ is a disc of radius r centered at the origin. Under CSR, for a process with intensity λ, $F(r) = 1 - \exp(-\lambda\pi r^2)$.

Replacing the location x in $F(r)$ by a point of the process, another summary statistic is obtained. This is called the *nearest neighbour distance distribution function*, and denoted by $G(r)$. The interpretation is that $G(r)$ is the probability that the distance from a randomly chosen point of the process to its nearest neighbour is less than or equal to r. Again, from stationarity, one can condition on having a point of the process at the origin, and express $G(r)$ as

$$G(r) = 1 - P_o(N_X(b(o,r)) = 0), \tag{1.5}$$

where $P_o(\cdot)$ is the conditional probability given that there is a point of the process at the origin. Note that the probability of the process having a point at the origin is zero, so the conditional probability cannot be defined in the classical way, by dividing by the probability of the event that is conditioned on. To properly define these conditional probabilities, one needs the concept of *Palm distributions*. A full introduction of this concept is outside the scope of this thesis, and the interested reader is referred to, e.g. Lieshout (2000).

Another important summary statistic is Ripley's K-function, denoted by $K(r)$. The intuitive definition is that if X is a stationary and isotropic point process, with intensity λ, then $\lambda K(r)$ gives the expected number of further points within distance r from a typical point of the process. A strength of this summary statistic is that it contains information about the structure of the point pattern over a wide range of distances, while the previously mentioned $G(r)$ and $F(r)$ are more "short sighted". The mathematical definition of $K(r)$ is given by

$$\lambda K(r) = \mathbb{E}_o[N_X(b(o,r) \setminus \{o\}], \tag{1.6}$$

where the expectation is with respect to the Palm distribution mentioned above. It is also worth noting that the K-function is scaled by the intensity of the process. For example, if X is a CSR process, then $K(r) = \pi r^2$, which does not depend on λ.

Edge corrections

The summary statistics mentioned above are defined in terms of a point process and when working with point pattern data one aim is typically to characterize the process that generated the data in terms of one or more of the summary statistics. Thus, estimators of the summary statistics are needed, but the naive estimators one would first think of are typically biased, due to edge effects. Here, $K(r)$ will serve to illustrate some of the techniques available to compensate for edge effects.

As mentioned, $\lambda K(r)$ is the expected number of further points within distance r from a typical point of the process, and the common procedure to estimate an expectation from data is to replace it with the corresponding mean. An estimate of $K(r)$ can then be obtained as $\hat{K}(r) = \hat{\lambda}^{-1}\bar{n}(r)$, where $\bar{n}(r)$ is the estimate of the expectation $\lambda K(r)$ and $\hat{\lambda}$ is an estimate of the intensity of the process. Typically, $\hat{\lambda} = (n-1)/\nu(W)$ is used, where n is the number of points in W and $\nu(W)$ is the Lebesgue measure of W, rather than the obvious choice $\hat{\lambda} = n/\nu(W)$, for technical reasons. The mean, $\bar{n}(r)$, in this setting can be written as

$$\bar{n}(r) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\mathbf{1}\{||x_i - x_j|| \leq r\}, \tag{1.7}$$

i.e. for each observed point the number of r-close neighbours is counted and the total count is divided by n. Unfortunately, this is a biased estimator of $\lambda K(r)$, and the reason for this is that the observed pattern is restricted to the window, W. It might be that a point, x_i, has neighbours that are within distance r but outside the observation window, and therefore not included in the sum (1.7). This introduces bias to $\bar{n}(r)$ in (1.7) as an estimator $\lambda K(r)$. If the information about the pattern outside W is not available, as with the data used in this thesis, a common way to obtain an unbiased estimate is to introduce weights for the terms in the double sum, i.e. letting

$$\bar{n}(r) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i} w(x_i, x_j)\mathbf{1}\{||x_i - x_j|| \leq r\}. \tag{1.8}$$

Depending on the nature of the underlying process, different choices of w can be used. If the underlying process is (assumed to be) stationary, the *translational* edge correction can be employed, where

$$w(x_i, x_j) = \frac{1}{\nu(W_{x_i} \cap W_{x_j})}, \tag{1.9}$$

where W_{x_i} is the observation window translated by x_i. In other words, the weight is inversely proportional to the area of the intersection of the observation window translated by x_i and x_j. If the process is also isotropic the *isotropic* edge correction can be used, where $1/w(x_i, x_j)$ is the proportion of the circumference of the circle centered at x_i and having radius $||x_i - x_j||$ contained in W. This can be expressed as

$$w(x_i, x_j) = \frac{2\pi ||x_i - x_j||}{\nu_1(\partial b(x_i, ||x_i - x_j||) \cap W)}. \tag{1.10}$$

where ν_1 is Lebesgue measure in one dimension.

A third example of an edge correction method is the *border correction*. With this method, for a given r, only points that are further away from the boundary than r are considered as reference points, while all points are considered when counting the number of neighbours for a reference point. Specifically, let n_r denote the number of points in $X \cap W$ that are at least at distance r from the boundary of W, and denote these points by $\{x_i'\}_{i=1}^{n_r}$. Furthermore, let $n_i(r)$ denote the number of r-close neighbours of x_i' among all the points in W. The mean can then be expressed as

$$\bar{n}(r) = \frac{1}{n_r} \sum_{i=1}^{n_r} n_i(r), \tag{1.11}$$

and $\bar{n}(r)$ can be used to estimate $K(r)$, as above, by using $\hat{K}(r) = \hat{\lambda}^{-1}\bar{n}(r)$.

1.3 Hierarchical models

The data we study come with a natural hierarchy to it since we have statistics on different levels of the data, with levels ranging from groups of subjects to nerve fiber segments. Thus we can summarize the data at these different levels. This structure of the data, and how it is used in the modelling, is explained in the sections below.

Hierarchy in the data

The top level of the hierarchy is the groups of subjects, at which we can mainly use averages as observations. The next level is the subject level, and from each subject we have data from several blisters. At the subject level we have, for example, covariates, such as age and BMI as observations. The blister level is not considered in the modelling for reasons explained below. From each blister we have several samples from which the number of nerve fibers is observed. The next level is the nerve fibers, for which we use the sum of the segment lengths and maximum order of segments as observations, with the order of a segment being the number of times the fiber has branched before this segment. Finally, the lowest level is the segment level, where we consider the lengths and orders as observations.

When modelling the data, it is possible to disregard some of the levels in the structure. For example, in our analyses so far, we have found very little indication that samples from the same blister of one subject are more similar than samples from different blisters of the same subject. Therefore, the blister level is not considered, but all sample-level observations from one subject are assumed to have the same distribution. This observation is important in itself, since it suggests that the nerve fiber pattern of a subject is quite constant on this spatial scale, leading us to consider all samples from one subject as replicated observations of the same process.

Another thing to note, important when attempting both supervised and unsupervised classification on this data, is that we may have varying number of observations at each level for different subjects. The standard classification techniques, e.g. linear discriminant analysis and random forest classification for supervised classification and cluster analysis for unsupervised classification, are based on considering a fixed number of variables, called *features*. To implement such a method on data with differing number of observations, one has to choose a way to summarize the observations into a fixed set of features. Using hierarchical models as the ones described below, feature design is not needed, and all observations can be used as they are.

We also impose a type of hierarchy within a level specifying the models by giving conditional distributions where some of the responses depend on others. For example, the length of a nerve fiber segment is modelled to depend on, among other things, the order of that segment. This is expressed as a conditional distribution of the length, given the explanatory variables for that response, one of which is the order.

In the modelling we have done so far, we have only made use of part of the hierarchical structure. Observations from the nerve fiber and segment levels are modelled as depending only on the group belonging of each subject from which the nerve fiber or segment is observed, while individual effects could be added for these observations, for example, for the sample and subject levels. Hierarchical models offer a rich framework for future research. The structure allows for modelling observations of different kinds simultaneously, and for capturing variability at different levels of the data.

Mixture models

The diabetic subjects in our data are diagnosed as suffering from mild, moderate or severe diabetic neuropathy based on nerve counts at six body locations and a neuropathy score, and using some other observations might lead to a different grouping of the subjects, revealing other aspects of the neuropathy. Part of the work in this thesis aims to find a grouping of the subjects reflecting the severity of the disorder, based on the observed nerve fiber structures, and the nerve fiber density. For this, mixture models are used, in the framework of the hierarchical models described above. A mixture model is based on the assumption that the observations in the data come from different classes, for which the distributions of the observations may differ. A general mixture model can be described as follows: let y denote the response variable, Z the class belonging and n the number of classes in the model. The density of y can then be expressed as

$$p(y) = \sum_{z=1}^{n} P(Z = z)p(y|Z = z), \tag{1.12}$$

where $p(y|Z = z)$ is the density of y for a sample from class z and $P(Z = z)$ is the probability that a sample is from class z.

One way to use this class of models is to do unsupervised classification, or clustering. This means that the class belongings are modelled as unobserved random variables, with Z_i being the class belonging of observation i. Fitting the model provides estimates of the parameters, $\hat{\theta}$, but also estimates of the probabilities of the class belongings given the parameter estimates, $\hat{w}_{iz} = P(Z_i = z|y_i, \hat{\theta})$. The latter estimates can be used to group the data into n classes, by classifying the observation y_i into $z^* = \arg\max_z(P(Z_i = z|y_i, \hat{\theta}))$. Perhaps the most popular method for fitting this type of models is the expectation maximization (EM) algorithm (Dempster et al., 1977). First, initial values of the parameters and the probabilities of class belongings, w_{iz}, are given. The algorithm then proceeds iteratively by updating the parameter values given the w_{iz}'s, and then updates the w_{iz}'s given the current parameter values.

As a simple example of this, consider the simulated data from a multivariate normal mixture distribution with two components, given in Figure 1.3, having means and covariance matrices

$$\mu_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \qquad \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0.75 \end{bmatrix}$$

$$\mu_2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 0.5 & -0.25 \\ -0.25 & 0.5 \end{bmatrix},$$

and proportions $P(Z = 1) = 0.4$ and $P(Z = 2) = 0.6$. This model is fitted to the data, and in Figure 1.4 is an illustration of the results showing the true class and fitted class of each data point. The model captures the two clusters well, and only a few data points are misclassified. However, this is an ideal case when the correct model with the correct number of classes is fitted to the data, while in practical applications, the appropriate distributions and number of classes to use are not always known.

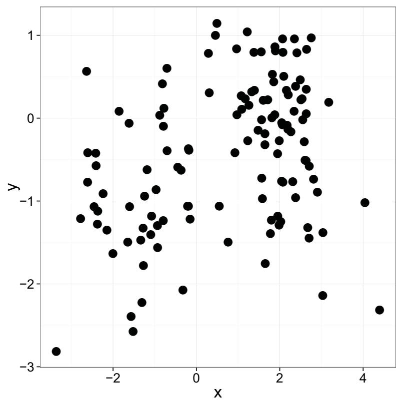

Figure 1.3: Simulated data from a multivariate normal mixture distribution.

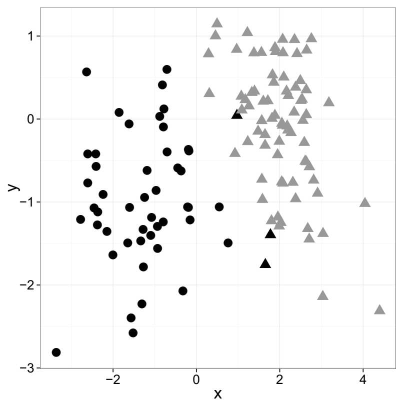

Figure 1.4: Illustration of the classification. The true classes are shown in different colors, and the fitted classes in different shapes. The black data points are from the class 1 and the gray data points from class 2, while the dots are classified as 1 and the triangles are classified as 2. Thus any gray circle or black triangle is a misclassification.

Chapter Two

References

Chiu, S. N., Stoyan, D., Kendall, W. S., and Mecke, J. (2013). *Stochastic geometry and its applications*. John Wiley & Sons: Chichester.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons: Chichester.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B*, pages 1–38.

Diggle, P. J. (2014). *Statistical analysis of spatial and spatio-temporal point patterns*. Chapman & Hall: Boca Raton.

Gelfand, A. E., Diggle, P., Fuentes, M., and Guttorp, P. (2010). *Handbook of spatial statistics*. CRC Press: Boca Raton.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*. John Wiley & Sons: Chichester.

Kennedy, W. R., Nolano, M., Wendelschafer-Crabb, G., Johnson, T. L., and Tamura, E. (1999). A skin blister method to study epidermal nerves in peripheral nerve disease. *Muscle & Nerve*, 22(3):360–371.

Kennedy, W. R. and Wendelschafer-Crabb, G. (1993). The innervation of human epidermis. *Journal of the neurological sciences*, 115(2):184–190.

Lauria, G., Bakkers, M., Schmitz, C., Lombardi, R., Penza, P., Devigili, G., Smith, A. G., Hsieh, S.-T., Mellgren, S. I., Umapathi, T., et al. (2010a). Intraepidermal nerve fiber density at the distal leg: a worldwide normative reference study. *Journal of the Peripheral Nervous System*, 15(3):202–207.

Lauria, G., Hsieh, S. T., Johansson, O., Kennedy, W. R., Leger, J. M., Mellgren, S. I., Nolano, M., Merkies, I. S., Polydefkis, M., Smith, A. G., et al. (2010b). European federation of neurological societies/peripheral nerve society guideline on the use of skin biopsy in the diagnosis of small fiber neuropathy. report of a joint task force of the european fe-deration of neurological societies and the peripheral nerve society. *European Journal of Neurology*, 17(7):903–e49.

Lieshout, v. M. (2000). *Markov point processes and their applications*. Imperial College Press: London.

Møller, J. and Waagepetersen, R. P. (2003). *Statistical inference and simulation for spatial point processes*. CRC Press: Boca Ranton.

Myllymäki, M., Panoutsopoulou, I., and Särkkä, A. (2012). Analysis of spatial structure of epidermal nerve entry point patterns based on replicated data. *Journal of Microscopy*, 247(3):228–239.

Myllymäki, M., Särkkä, A., and Vehtari, A. (2014). Hierarchical second-order analysis of replicated spatial point patterns with non-spatial covariates. *Spatial Statistics*, 8:104–121.

Ogata, Y. and Tanemura, M. (1985). Estimation of interaction potentials of marked spatial point patterns through the maximum likelihood method. *Biometrics*, 41(2):421–433.

Olsbo, V., Myllymäki, M., Waller, L. A., and Särkkä, A. (2013). Development and evaluation of spatial point process models for epidermal nerve fibers. *Mathematical Biosciences*, 243(2):178–189.

Takacs, R. and Fiksel, T. (1986). Interaction pair-potentials for a system of ant's nests. *Biometrical Journal*, 28(8):1007–1013.

Waller, L. A., Särkkä, A., Olsbo, V., Myllymäki, M., Panoutsopoulou, I. G., Kennedy, W. R., and Wendelschafer-Crabb, G. (2011). Second-order spatial analysis of epidermal nerve fibers. *Statistics in Medicine*, 30(23):2827–2841.